交差確認法を用いた予測誤差の推定について

伏木　忠義　　数理・推論研究系　助教

2010年7月9日　統計数理研究所　オープンハウス

1. はじめに

興味が予測にある場合には，データから構成したモデルを使って予測するときにどの程度の予測誤差があるかを知っておくことは重要なことである．本発表では予測誤差の推定方法について議論する．

2. 問題設定

・データ：

$$\mathcal{D} = \{(x_1, y_1), \cdots, (x_N, y_N)\} \sim F \quad \text{i.i.d.}$$

・x が与えられたときに $\{h(x;\theta)\}$ という関数を使って y を予測．

・推定量：

$$\hat{\theta}(\mathcal{D}) = \operatorname*{argmin}_{\theta} \left[\sum_{i=1}^{N} \{y_i - h(x_i;\theta)\}^2 \right].$$

・予測誤差：

$$s_N = \mathrm{E}_{(x,y)\sim F}\{y - h(x;\hat{\theta}(\mathcal{D}))\}^2.$$

3. 訓練誤差と K-fold cross-validation

・訓練誤差：

$$\mathrm{TR}_N = \frac{1}{N}\sum_{i=1}^{N}\{y_i - h(x_i;\hat{\theta}(\mathcal{D}))\}^2.$$

・データをほぼ等分割：$\mathcal{D} = \{\mathcal{D}^{(1)}, \cdots, \mathcal{D}^{(K)}\}, \quad \mathcal{D}^{(-\alpha)} = \mathcal{D} \setminus \mathcal{D}^{(\alpha)}$.

・K-fold cross-validation：

$$\mathrm{CV}_{N,K} = N^{-1}\sum_{\alpha=1}^{K}\sum_{(x,y)\in\mathcal{D}^{(\alpha)}}\{y - h(x;\hat{\theta}(\mathcal{D}^{(-\alpha)}))\}^2.$$

・バイアス：

$$\mathrm{E}(\mathrm{TR}_N - s_N) < 0, \quad \mathrm{E}(\mathrm{CV}_{N,K} - s_N) > 0.$$

4. 訓練誤差と K-fold cross-validation をつなげる族

1. $\{\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda) | 0 \le \lambda \le 1\}$,

$$\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda) = (1-\lambda)\mathrm{CV}_{N,K} + \lambda\mathrm{TR}_N.$$

2. $\{\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda) | 0 \le \lambda \le 1\}$,

$$\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda) = N^{-1}\sum_{\alpha=1}^{K}\sum_{(x,y)\in\mathcal{D}^{(\alpha)}}\{y - h(x;\hat{\theta}(\mathcal{D}^{\lambda(-\alpha)}))\}^2.$$

ここで

$$\hat{\theta}(\mathcal{D}^{\lambda(-\alpha)}) = \operatorname*{argmin}_{\theta}\left[\sum_{(x,y)\in D^{(-\alpha)}}\{y - h(x;\theta)\}^2 + \lambda\sum_{(x,y)\in D^{(\alpha)}}\{y - h(x;\theta)\}^2\right].$$

($\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda)$ と $\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda)$ に関する注意)

・$\lambda = 0$ のとき：K-fold cross-validation $\mathrm{CV}_{N,K}$ に一致．

・$\lambda = 1$ のとき：訓練誤差 TR_N に一致．

・$K = N, \lambda = 1/(2N) + o(N^{-2})$ のとき：Yanagihara et al. (2006) で考えているバイアス補正された cross-validation と一致．

5. バイアス補正された K-fold cross-validation

1. $\lambda^{\mathrm{M}} = (2K-1)^{-1}$ とする．このとき，$\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda^{\mathrm{M}})$ は漸近的にバイアス補正されている．

2. $\lambda^{\mathrm{E}} = (K-1)\{(1-K^{-2})^{-1/2} - 1\}$ とする．このとき，$\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda^{\mathrm{E}})$ は漸近的にバイアス補正されている．

(注意)

・上の λ^{M}，λ^{E} は K だけの関数として書けているのでさらなる推定をしなくてもこの方法を適用することができる．

・$K = N$ のとき，$\lambda^{\mathrm{M}} = 1/(2N) + o(N^{-2})$，$\lambda^{\mathrm{E}} = 1/(2N) + o(N^{-2})$.

6. 例

1. 線形回帰.

・真の分布：

$$Y = \beta_0 + \sum_{k=1}^{d-1}\beta_k X_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma_0^2), \quad X \sim U(-1,1)^{d-1}$$

線形回帰モデルで最小二乗法を使って β を推定する．

・正規化された予測誤差：

$$s_N = \mathrm{E}_{(X,Y)}[(Y - \hat{\beta}^T X)^2]/(2\sigma_0^2).$$

・$N = 1000, d = 250$ のときの結果：

	$K=5$	$K=10$	$K=N$
$\mathrm{E}_{\mathcal{D}}[\mathrm{CV}_{N,K}]$	0.727	0.693	**0.667**
$\mathrm{E}_{\mathcal{D}}[\mathrm{CV}^{\mathrm{M}}_{N,K}]$	0.688	0.676	**0.667**
$\mathrm{E}_{\mathcal{D}}[\mathrm{CV}^{\mathrm{E}}_{N,K}]$	**0.662**	**0.666**	**0.667**
$\mathrm{s.d.}_{\mathcal{D}}[\mathrm{CV}_{N,K} - s_N]$	0.046	0.042	**0.039**
$\mathrm{s.d.}_{\mathcal{D}}[\mathrm{CV}^{\mathrm{M}}_{N,K} - s_N]$	0.043	**0.040**	**0.039**
$\mathrm{s.d.}_{\mathcal{D}}[\mathrm{CV}^{\mathrm{E}}_{N,K} - s_N]$	**0.041**	**0.040**	**0.039**

・$\mathrm{E}_{\mathcal{D}}[s_N] = 0.667$.

・真値 β は $U(-1,1)^d$ を使って発生．

・$\mathcal{D}$ の期待値は10000回の Monte Carlo 計算によって求めた．

2. 非線形回帰.

・真の分布：

$$\begin{aligned} Y &= f_0(X) + \varepsilon, \quad \varepsilon \sim N(0,\sigma_0^2),\ X \sim N(0,1)^{d-1}, \\ f_0(X) &= 1 - 3\varsigma(1 + X_1 + 3X_2 - X_3 - 2X_4 + 5X_5) \\ &\quad +5\varsigma(-2 + 2X_1 - 3X_2 + X_3 + 2X_4), \\ \varsigma(x) &= \{1+\exp(-x)\}^{-1} \end{aligned}$$

非線形回帰モデル $h(x;\theta) = \beta_0 + \sum\beta_i\varsigma(w_{i0} + w_i^T X)$ を仮定し最小二乗法を使って θ を推定．

・正規化された予測誤差：

$$s_N = \mathrm{E}_{(X,Y)}[\{Y - h(X;\hat{\theta}(\mathcal{D})\}^2]/(2\sigma_0^2).$$

・$N = 80$ のときの結果：

	$K=5$	$K=10$	$K=N$
$\mathrm{E}_{\mathcal{D}}[\mathrm{CV}_{N,K}]$	0.793	0.726	0.688
$\mathrm{E}_{\mathcal{D}}[\mathrm{CV}^{\mathrm{M}}_{N,K}]$	0.750	0.709	0.686
$\mathrm{E}_{\mathcal{D}}[\mathrm{CV}^{\mathrm{E}}_{N,K}]$	**0.671**	**0.679**	**0.681**
$\mathrm{s.d.}_{\mathcal{D}}[\mathrm{CV}_{N,K} - s_N]$	0.243	0.268	0.195
$\mathrm{s.d.}_{\mathcal{D}}[\mathrm{CV}^{\mathrm{M}}_{N,K} - s_N]$	0.225	0.260	0.194
$\mathrm{s.d.}_{\mathcal{D}}[\mathrm{CV}^{\mathrm{E}}_{N,K} - s_N]$	**0.194**	**0.218**	**0.187**

・$\mathrm{E}_{\mathcal{D}}[s_N] = 0.681$.

・$\mathcal{D}$ の期待値は10000回の Monte Carlo 計算によって求めた．

参考文献

Fushiki, T. Estimation of prediction error by using K-fold cross-validation, *Statistics and Computing*, (to appear).

Yanagihara, H., Tonda, T., & Matsumoto, C. Bias correction of cross-validation criterion based on Kullback-Leibler information under a general condition. *Journal of Multivariate Analysis*, **97**, 1965-1975. (2006).